



WORTSCHATZ
LEIPZIG

Working with (English News) Corpora at the Leipzig Corpora Collection

Thomas Eckart (eckart@saw-leipzig.de)
Felix Helfer (helfer@saw-leipzig.de)
Erik Körner (koerner@saw-leipzig.de)



Project Wortschatz Leipzig



Wortschatz Leipzig

Short project history

- Started in 1994, first Web portal in 1998, first API since 2004
- Text databases with more than 30 billion sentences in more than 250 languages
- Strong focus on “Web as a corpus” paradigm
- Cooperation between Leipzig University, SAW and InfAI

Competences

- Handling large amounts of textual content
- Investing in algorithms instead of individual data sets
- Development of procedures independent of individual languages to internationalize with minimal effort



NFDI – Text+

Participating in the Text+ project via the *Saxon Academy of Sciences and Humanities*

Building on long experience in providing linguistic resources to users in

- Data domain “lexical resources”
 - Providing our own dictionaries in the NFDI
 - Working on interfaces and search engines for accessing lexical resources
- Task area “infrastructure / operations”
 - Findability & Interoperability of text-based resources in a federated infrastructure
 - *Federated Content Search* (FCS) as central search component in Text+
 - (Linguistic) Linked Open Data



Wortschatz Leipzig – Thematic focus

Text analysis

- Compiling large corpora as text databases
- Use of a complex preprocessing pipeline
- Focus on statistical analysis:
 - Frequency- and cooccurrences-based analysis, sentiment analysis, topic modelling, document classification

Applications

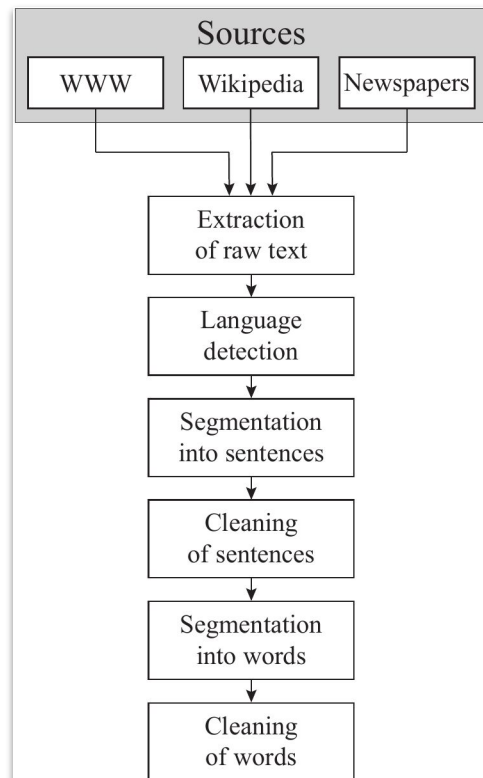
- Support with dictionary work and ontology creation
- Knowledge management / information retrieval: extraction of connections & keywords etc.



Wortschatz Leipzig – Sources & Preprocessing

- Online news sites
 - RSS-Feeds (daily, in ca. 80 languages)
- Web text
 - Random crawling (of selected TLDs)
 - Stopword-driven (BootCat)
 - User-driven crawling
- Wikipedia
- Administrative texts, ...

- Pipeline for crawling, cleaning, language-identification, statistical analysis, publication



Web data

Crawling since 2002; over a trillion crawled HTML-pages so far.

250+ languages available online - for an overview, see the language selector on:

<https://corpora.wortschatz-leipzig.de/en>



Example: English news corpus 2023*

- Based on daily RSS feeds
- Around 3,7M documents from 3400 news sites in English language all over the globe (66% .com, 17% .uk, 3% .au, 2,8% .net,...)
- Around 70M sentences remaining after standard toolchain processing and cleaning
- Annotations:
 - POS tagged
 - POS tag based sentence signatures
 - Word cooccurrences (windows: sentence & ± 1 token)
 - Cooccurrences-based word embeddings
 - ...

* up to 2023-11-19



Applications



Corpus portal



- Idea: distant reading approach to corpora, condensed presentation of linguistic and statistical analysis
- References to other WS portals
- Only selection of available data

The screenshot shows the Text+ corpus portal interface. At the top, there are dropdown menus for selecting a language. The 'English' language is selected. Below the language selection, there are search results for 'NEWS' and 'NEWS-ECONOMY'. The 'NEWS' section lists various news items from 2008 to 2023. The 'NEWS-ECONOMY' section lists news items from 2002 to 2023. The 'WEB' section lists various web pages from 2002 to 2019. The 'WIKIPEDIA' section lists Wikipedia articles from 2021. At the bottom, there are more language selection options.

NEWS	NEWS-ECONOMY	WEB	WIKIPEDIA
News 2008	News-economy 2023	Web 2018 (Thailand)	
News 2012		Web 2019 (Cambodia)	
News 2012		Web 2019 (Ireland)	
News 2013		Web 2019 (Namibia)	
News 2016		Web 2019 (South Africa)	
News 2016 (South Africa)		Web 2019 (Thailand)	
News 2017 (South Africa)		Web 2019 (United Kingdom of Great B)	
News 2018		Web 2019 (Zambia)	
News 2019		Web 2019 (Zimbabwe)	
News 2020			Wikipedia 2021
✓ News 2023			Wikipedia 2021

→ <https://corpora.wortschatz-leipzig.de/en>

<https://www.text-plus.org>



Corpus portal – German News 2022 – *Universität*

Frequency

Linguistic information

Definitions

Dornseiff

Embedding similarity

Examples

[...]

The screenshot shows the Corpus portal interface for the word "Universität". The top bar displays the search term "Universität" and corpus metadata: "German news corpus based on material from 2022 with 31,774,802 sentences." The main content area is divided into two columns. The left column contains linguistic and definitional information: "Word: Universität" with 29,312 occurrences, rank 1,598, and frequency class 9. It lists "See also: Universität", "Article: die", "Part of speech: Noun", "Baseform of: Universitäten", and "Abbreviation: UNI". The "Description" section includes bullet points about universities and a synonym list: "Fachakademie, College, Forschungsanstalt, Hochschule, Uni, Lehranstalt, Akademie". Below this are sections for "Dornseiff Sets", "13.1 Studium, Universität" (listing various types of universities), "Words with Similar Context" (listing "Uni", "Die Universität", "Universitäten", "Hochschule" with similarity scores), and "Examples" (with a sample sentence). The right column features a "Word graph" showing a network of related terms like "Universität E", "Universität Wien", "Innsbruck", "Studie", "Technischen", "Die Universität", "Professor", "Universität Graz", "Universität Innsbruck", "Wien", and "Universität M", "Universität Zürich".

Corpus metadata & Links

Cooccurrences graph

→ https://corpora.wortschatz-leipzig.de/en?corpusId=deu_news_2022&word=Universität



Corpus portal – English News 2023 – values

Frequency

Various linguistic information

String similarity

Embedding similarity

Examples

[Word co-occurrences]

The screenshot shows the Corpus Portal interface for the word "values". At the top, it indicates the corpus is based on material from 2023 with 70,488,310 sentences. The main section displays the word "values" with its frequency (64,096), rank (2,597), and frequency class (10). It also lists linguistic information such as "Part of speech: Verb, Noun" and "Baseform: value". A section titled "Words with Similar Context" lists related terms like "principles", "ideals", "beliefs", etc., with their respective scores. Below this, "Examples" are provided, showing sentences from news articles where "values" is used. On the right side, a "Word graph" is displayed, showing a network of related terms connected by lines, with "values" at the center. The graph includes terms like "American values", "property", "shared", "culture", "core", "principles", "democratic", "beliefs", "align with", "respect", "Christian values", and "align".

Corpus metadata & Links

Cooccurrences graph

→ https://corpora.wortschatz-leipzig.de/en?corpusId=eng_news_2023&word=values



Corpus portal – Subcorpus “economy” – values

- Filtered subcorpus, part of *English news 2023*
- Document filter based on simple keyword extraction
- Size:
 - 197K documents
 - 5.2M sentences
 - 111M word tokens

The screenshot shows the Text+ corpus portal interface for the word "values". The search bar at the top contains "values". The page title is "English news corpus based on material from 2023 with 5,196,264 sentences".

Word: values Number of occurrences: 5,580 Rank: 2,135 Frequency class: 10

See also: Values
Part of speech: Verb, Noun
Baseform: value

Part of: Market values, American values, Christian values, Core values, Western values, target values, European values, British values, Jewish values, Judeo-Christian values, system of values

Words with Similar Context

goals (0.19), valuations (0.19), prices (0.19), objectives (0.19), culture (0.18), priorities (0.18), levels (0.16), value (0.16), interests (0.16), mission (0.16), principles (0.16), relationships (0.15), purpose (0.15), rents (0.15), yields (0.15), investments (0.14), standards (0.14), metrics (0.14), practices (0.14), owners (0.14), stories (0.14), volumes (0.14), strategy (0.14), experiences (0.14), vision (0.13), rates (0.13), assets (0.13), costs (0.13), outcomes (0.13), commitment (0.13), properties (0.13), partners (0.13), sales (0.13), communities (0.12), margins (0.12), transactions (0.12), issues (0.12), dynamics (0.12), loans (0.12), initiatives (0.12), homes (0.12), actions (0.12), stock price (0.12), returns (0.12), trends (0.12), profits (0.12), lives (0.12), portfolios (0.12), skills (0.12), clients (0.12), markets (0.12), businesses (0.12), numbers (0.11), efforts (0.11), purchases (0.11), conditions (0.11), categories (0.11), trust (0.11), brands (0.11), benefits (0.11)

Examples

- Assess which of these factors currently affects you and identify any misalignment with your **values**. (www.forbes.com, collected on 16/06/2023)
- But going back to Nick's question, that may show up in net asset **values** and property pricing. (seekingalpha.com, collected on 11/02/2023)
- Define clear and meaningful goals that align with your **values** and aspirations. (www.thenationalherald.com, collected on 14/05/2023)
- Mutual funds are not guaranteed, their **values** change frequently and past performance may not be repeated. (financialpost.com, collected on 27/01/2023)
- Spence is a founder who's obsessed with creating a culture that **values** creativity, curiosity and empathy. (www.bandt.com.au, collected on 03/08/2023)
- Mutual funds are not guaranteed, their **values** change frequently, and past performance may not be repeated. (www.cbj.ca, collected on 30/03/2023)
- Avoiding burnout starts with anchoring a career portfolio based on **values** and an ideal work life. (fortune.com, collected on 03/01/2023)
- In markets where property **values** have soared over the past decade, you will likely find yourself undercovered

Word graph

The word graph shows "values" at the center, connected to various related terms: home, property, align with, asset, align, core, culture, shared, Estimation, and Market values.

→ https://corpora.wortschatz-leipzig.de/en?corpusId=eng_news-economy_2023&word=values



“Words of the day”

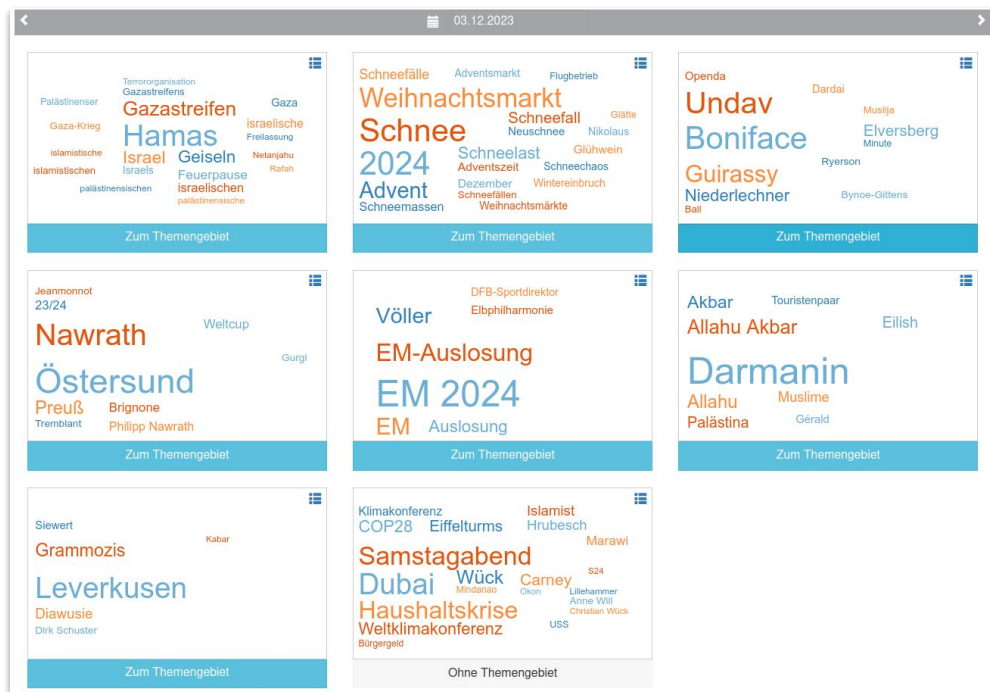
- Idea: daily news text can be used to compare thematic developments and identify salient topics on a daily basis
- Topic Modelling (based on LDA using a standard library)
- Material:
 - German: ca. 8,000 articles per day (75% .de, 15% .at, 4% .ch)
 - English: ca. 16,000 articles per day (worldwide)
- Presentation at “words of the day” portal (currently only in German language)

→ <https://wod.wortschatz-leipzig.de/>



Topics of the day

2023-12-03



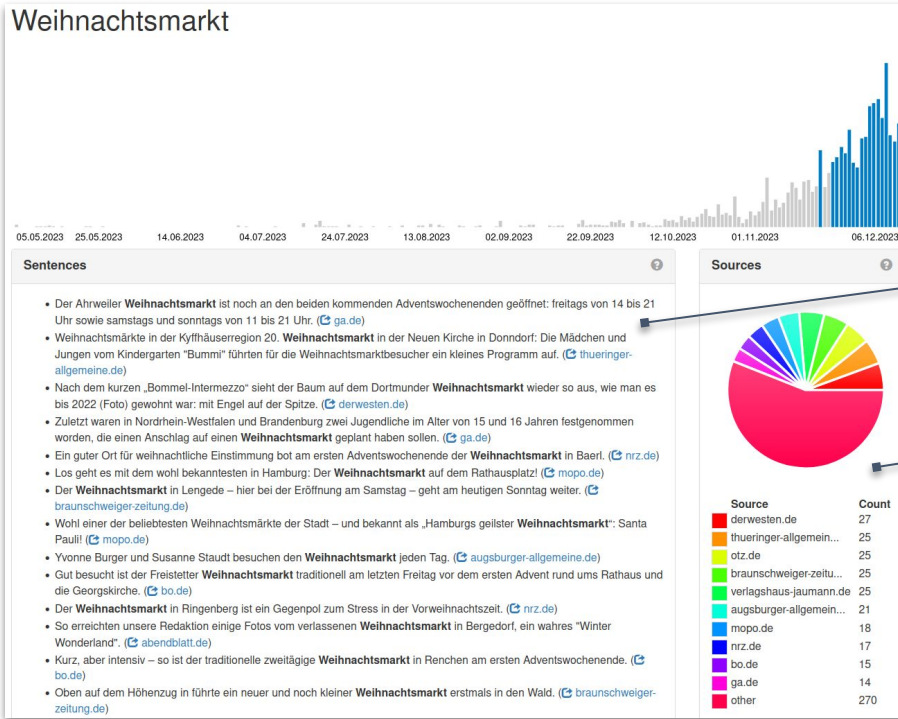
→ <https://wod.wortschatz-leipzig.de/>

<https://www.text-plus.org>



Diachronic analysis of words – *Christmas market*

Weihnachtsmarkt



Relative frequency over time

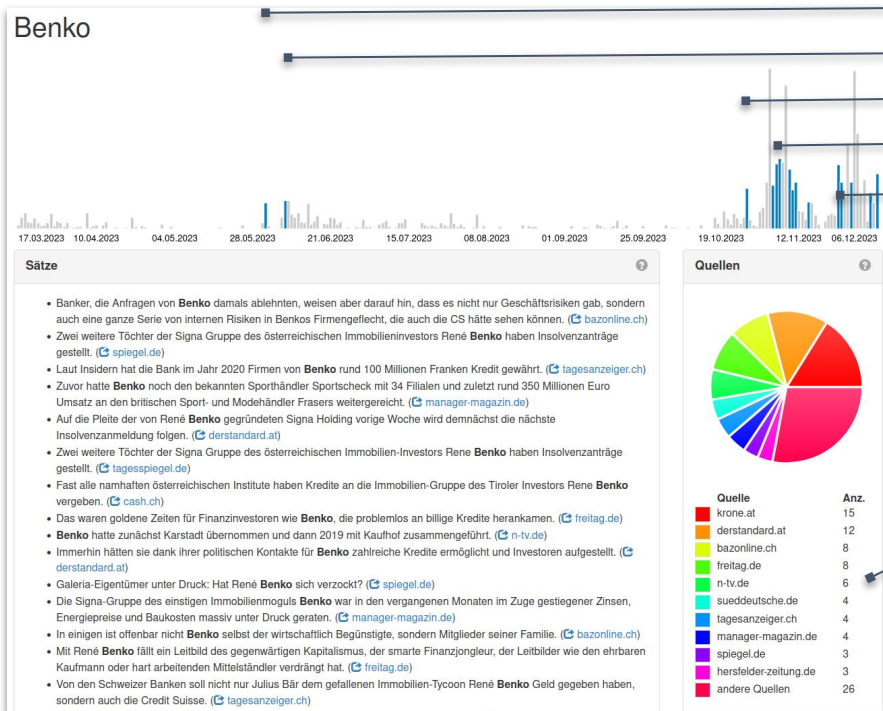
References

Sources

→ [Weihnachtsmarkt](#)



Diachronic analysis of words – Benko



- *Signa Holding sells kika/Leiner*
- *Insolvency of kika/Leiner*
- *Insolvency of Signa Sports United*
- *Leaves Signa Holding*
- *Insolvency of Signa Holding*

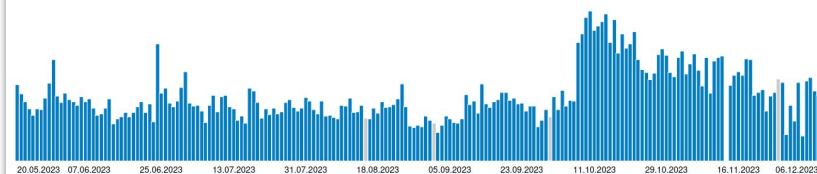
Quelle	Anz.
krone.at	15
derstandard.at	12
bazonline.ch	8
freitag.de	8
n-tv.de	6
sueddeutsche.de	4
tagesanzeiger.ch	4
manager-magazin.de	4
spiegel.de	3
herstelder-zeitung.de	3
andere Quellen	26

→ [Benko](#)



Diachronic development of topics – topic “war”

Topic: Hamas, Gazastreifen, Israel, ...



Words

Hamas • Gazastreifen • Israel • Geiseln • Feuerpause • israelischen • israelische • Gaza • Israels • islamistischen • Gaza-Krieg • Gazastreifens • Palästinenser • islamistische • palästinensische • palästinensischen • Terrororganisation • Netanjahu • Rafah • Freilassung • Kampfpause • freigelassen • abgeriegelten • Gaza-Streifen • Junis • Israels • israelischer • Benjamin Netanjahu • Huthi-Rebellen • Gaza-Kriegs • Terroristen • Terrorziele • Häftlinge • Hamas-Terroristen • Chan • Bodenoffensive • Küstenstreifens • Küstengebiets • DOSB • IDF • Palästinensergebiet • freigelassenen • freikommen • Mete • Joschka Fischer • Joschka • Nahost

Sources

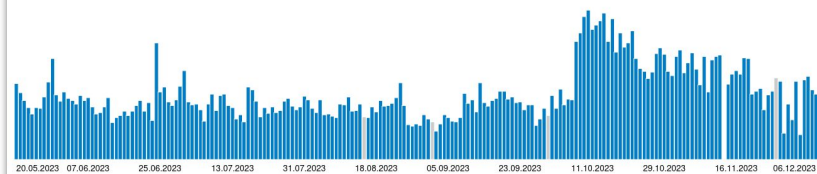


Source	Count
sueddeutsche.de	32
n-tv.de	31
faz.net	29
zeit.de	27
welt.de	27
bild.de	25
stern.de	22
derstandard.at	21
fr.de	21
spiegel.de	17
other	501

Sentences

- Die **israelische** Armee hat nach eigenen Angaben seit Beginn des **Gaza**-Kriegs mehr als 800 Tunnelschächte gelunden. (ad-hoc-news.de)
- Er habe gegenüber Herzog wiederholt, was er vergangene Woche in **Rafah** gesagt habe: „Keine Tötung von Zivilisten mehr“. (tagesschau.de)
- Lehrer warf den Organisatoren vor, das Selbstverteidigungsrecht **Israels** zu leugnen. (azonline.de)
- Premierminister **Benjamin Netanjahu**, hier beim Besuch von Joe Biden Mitte Oktober, widersetzt sich den Forderungen aus Washington, die Kriegsführung zu ändern. (sueddeutsche.de)
- Augenzeugen hatten der Deutschen Presse-Agentur am Sonntag berichtet, **israelische** seien in ein Gebiet östlich der Stadt **Chan Junis** im Süden des **Gazastreifens** vorgerückt. (stern.de)
- Wie viele **Geiseln** am Sonntag **freikommen** könnten, wurde nicht mitgeteilt. (braunschweiger-zeitung.de)
- Die **israelischen** Streitkräfte gaben am Sonntag an, seit Beginn des Gazakriegs rund 10.000 Luftangriffe auf Ziele in dem **abgeriegelten Palästinensergebiet** durchgeführt zu haben. (bvz.at)
- Das Mädchen, ihre Zwillingsschwester sowie deren Mutter waren kürzlich im Rahmen eines Abkommens zwischen **Israel** und der **Hamas freigelassen** worden. (luzernerzeitung.ch)
- In der Zeit liess die **Hamas 105 Geiseln** frei, darunter 14 Deutsche, und **Israel** im Gegenzug 240 **palästinensische Häftlinge**. (luzernerzeitung.ch)

Topic: Belgorod, Prigoschin, Bachmut, ...



Words

Belgorod • Prigoschin • Bachmut • Freiwilligenkorps • Switolina • Chodakowski • Orcas • Gegenoffensive • Feldkommandeur • Gladkow

Sources



Source	Count
extremnews.com	58
kn-online.de	35
welt.de	34
tagesspiegel.de	29
n-tv.de	26
merkur.de	24
fr.de	21
faz.net	21
handelsblatt.com	20
spiegel.de	19
other	585

Sentences

- Statt von **Bachmut** aus weitere Angriffe auf benachbarte ostukrainische Gebiete zu starten, zogen sich die geschwächten Kämpfer ins bereits besetzte Hinterland zurück, heißt es laut unter Kriegs-Experten. (merkur.de)
- Grund dafür sei, dass Gouverneur **Gladkow** nicht zu einem geplanten Treffen mit den russischen Partisanen erschienen sei. (spiegel.de)
- Während es im Ukraine-Krieg nicht vorangeht, blickt Russland mit Sorge auf die drohende **Gegenoffensive** der ukrainischen Armee. (merkur.de)
- Russland hatte seinen Krieg gegen die am 24. Februar 2022 auch vom Gebiet **Belgorod** aus begonnen. (rp-online.de)
- Ein russischer **Feldkommandeur** sieht dies allerdings ganz anders. (n-tv.de)
- Bei den Eindringlingen scheint es sich jedenfalls um Mitglieder des bereits erwähnten "Russischen **Freiwilligenkorps**" zu handeln, das zwar aufseiten der Ukrainer kämpft, aber aus russischen Nationalisten besteht. (wienerzeitung.at)
- Die Gegend um **Bachmut** bleibe das „Epizentrum“ der Kämpfe. (welt.de)
- Update vom 5. Juni, 20.44 Uhr: Kämpfer eines **Freiwilligenkorps** bestehend aus russischen Nationalisten haben offenbar die Ortschaft Nowaja Tawischanka in der russischen Region **Belgorod** komplett unter ihre Kontrolle gebracht. (herfelder-zeitung.de)
- Konkret handelt es sich um ein polnisches **Freiwilligenkorps**, welches den Namen PDK trägt. (fr.de)
- Es erinnert an den Roman „Der Schwarm“ von Frank Schätzing: In Spanien häufen sich die Angriffe von **Orcas** auf Segelboote. (berliner-kurier.de)
- Bei den blutigen Kämpfen um die ostukrainische Stadt **Bachmut** soll es zu militärischen Auseinandersetzungen zwischen beiden Parteien gekommen sein. (merkur.de)

→ [2023-12-03](https://www.text-plus.org/2023-12-03)

→ [2023-06-05](https://www.text-plus.org/2023-06-05)



Corpus query engine NSE

- Idea: computer linguistics has long-term experience in working with large text collections
- Various tools to support work (filtering, selection, terminology work, sub-corpora comparisons etc.)
- One prominent application: NoSketchEngine (NSE)
 - Open-source version of a commercial tool, that focuses on lexicographers (“word sketches”)
 - Very efficient indexing of **large** text corpora
 - Allows: Filterung, dynamic sub-corpora generation, keyword extraction, corpus comparison, flexible word collocation presentation → distant **and** close reading approaches

→ <https://cql.wortschatz-leipzig.de/>






Corpus query engine NSE

Corpus Name	Language	Word Count	Action
abk_wikipedia_2012	Abkhazian	17,708	OPEN
abk_wikipedia_2021	Abkhazian	49,098	OPEN
ace_wikipedia_2021	Achinese	100,432	OPEN
ach_newscrawl_2011	Acoli	4,952	OPEN
afr_mixed_2013	Afrikaans	85,078,293	OPEN
afr_mixed_2014	Afrikaans	159,354,667	OPEN
afr_mixed_2019	Afrikaans	287,768,130	OPEN
afr_news_2019	Afrikaans	350,985	OPEN
afr_news_2020	Afrikaans	599,225	OPEN
afr_wikipedia_2018_300K	Afrikaans	5,429,842	OPEN

→ <https://cql.wortschatz-leipzig.de/#corpus?tab=advanced>



Corpus query engine NSE

- Example with extensive linguistic annotation: “eng_news_2023_topic”
- Processing of RSS/Feed based daily news crawl dataset with
 - **Document: LCC** html-to-text extraction and document language filtering
 - **Document: LCC** Sentence Segmentation and Tokenization
 - **Sentence: TextBlob**  Sentiment Polarity and Subjectivity prediction
 - **Sentence: Transformers**  AGNews topic classification
 - **Token: spaCy**  for linguistic annotation, e.g., lemmatization, *part-of-speech tagging, morphological features*, word shape, named entity recognition
- Statistics:
 - 5 Mio. documents, 417 Mio. sentences, 5.9 Bn. tokens

→ https://cql.wortschatz-leipzig.de/#dashboard?corpname=eng_news_2023_topic&corp_info=1
https://cql.wortschatz-leipzig.de/#dashboard?corpname=eng_news_2023_topic2&corp_info=1



Subcorpus definition and creation

Subcorpus definition

Subcorpus: **workdays**

Strukturattribut: <doc>

- doc.dayOfWeek = "Friday"
- doc.dayOfWeek = "Monday"
- doc.dayOfWeek = "Thursday"
- doc.dayOfWeek = "Tuesday"
- doc.dayOfWeek = "Wednesday"

SCHLIESSEN

Subkorpora based on text types or text patterns (CQL)

Name	business	business_high	not_business	weekend	workdays
					3.933.978.462 ~3.401.725.243

23,7
18,3
76,3
33
67

SubKORPUS ERSTELLEN

Show preloaded subcorpora

ZURÜCK

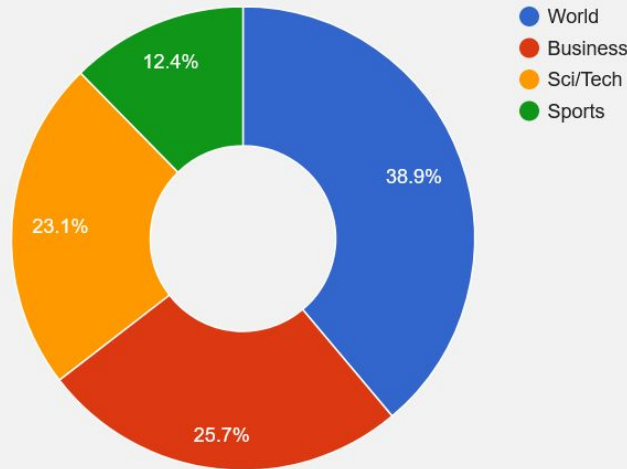
Please be aware that some of the presented sub-corpora and features are *currently only available via login* – Feel free to contact us!

→ https://cql.wortschatz-leipzig.de/#ca-subcorpora?corpname=eng_news_2023_topic



Distribution of sentences by AGNews topic

s - Topic classification using AGNews topics: World, Sports, Business, Sci/Tech



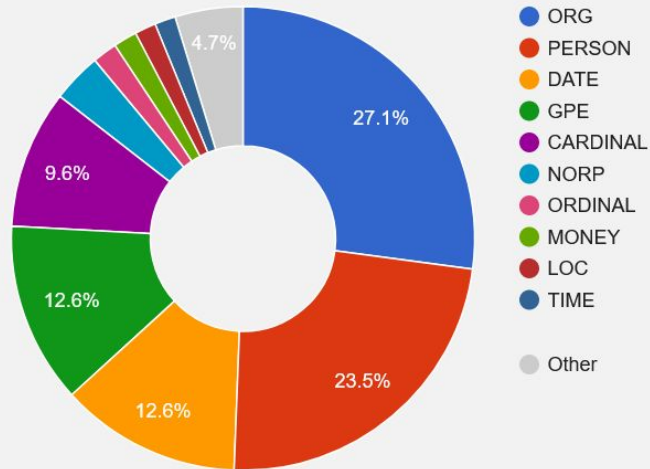
Attributwert	Structure frequency [?]
1 World	161.891.852 ...
2 Business	107.013.019 ...
3 Sci/Tech	96.120.559 ...
4 Sports	51.545.339 ...

→ [https://cql.wortschatz-leipzig.de/#text-type-analysis?corpname=eng_news_2023_topic&tab=basic&filter=containing&wlattrib=s.topic ...](https://cql.wortschatz-leipzig.de/#text-type-analysis?corpname=eng_news_2023_topic&tab=basic&filter=containing&wlattrib=s.topic...)



Distribution of Named Entities

named_entity - type



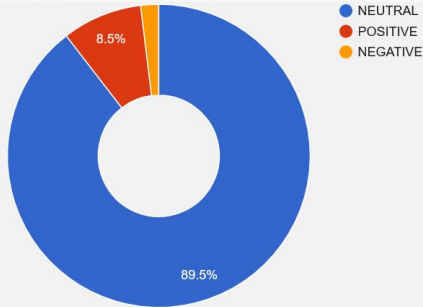
Attributwert	Structure frequency ?	Attributwert	Structure frequency ?
1 ORG	161.526.665 ...	11 FAC	7.295.712 ...
2 PERSON	140.127.419 ...	12 PRODUCT	6.622.137 ...
3 DATE	75.246.554 ...	13 WORK_OF_ART	6.195.678 ...
4 GPE	75.192.146 ...	14 EVENT	3.920.734 ...
5 CARDINAL	57.477.561 ...	15 QUANTITY	1.890.060 ...
6 NORP	20.556.376 ...	16 LAW	1.226.776 ...
7 ORDINAL	10.293.511 ...	17 PERCENT	636.034 ...
8 MONEY	9.805.882 ...	18 LANGUAGE	452.398 ...
9 LOC	8.847.620 ...		
10 TIME	8.700.431 ...		

→ https://cql.wortschatz-leipzig.de/#text-type-analysis?corpname=eng_news_2023_topic&tab=basic&filter=containing&onecolumn=1&wlatr=named_entity.type

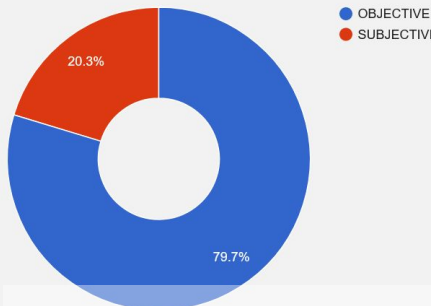


Frequency of Sentiment Polarity by Topic

s - TextBlob sentiment polarity as class: NEGATIVE (< 0.4), POSITIVE (> 0.4), NEUTRAL



s - TextBlob sentiment subjectivity as class: OBJECTIVE (< 0.5), SUBJECTIVE



Topic classification using AGNews topics	TextBlob sentiment polarity	Frequenz
1 Business	NEGATIVE	4.892 ...
2 Business	NEUTRAL	1.339.940 ...
3 Business	POSITIVE	94.274 ...
4 Sci/Tech	NEGATIVE	19.387 ...
5 Sci/Tech	NEUTRAL	1.604.824 ...
6 Sci/Tech	POSITIVE	74.020 ...
7 Sports	NEGATIVE	4.916 ...
8 Sports	NEUTRAL	344.074 ...
9 Sports	POSITIVE	14.612 ...
10 World	NEGATIVE	12.307 ...
11 World	NEUTRAL	1.459.671 ...
12 World	POSITIVE	82.111 ...

→ https://cql.wortschatz-leipzig.de/#concordance?corpname=eng_news_2023_topic&queryselector=cql&attrs=word&viewmode=kwic&attr_allpos=all&refs_up=0&...



Automatic keyword extraction

- Comparing word frequencies using a reference corpus (or subcorpus, e.g., business against non-business)

referenzkorpus: eng_news_2023_topic subkorpus: not_business (Wortanzahl: 803.891)

Word	Word	Word	Word	Word
1 gannett ...	11 withdrawn ...	21 buckinghamshire ...	31 dock ...	41 profits ...
2 np20 ...	12 securities ...	22 earnings ...	32 equity ...	42 insurance ...
3 chartist ...	13 wycombe ...	23 owns ...	33 corp ...	43 annualized ...
4 dissatisfied ...	14 stock ...	24 groupdo ...	34 newport ...	44 quarterly ...
5 misused ...	15 inflation ...	25 recession ...	35 abused ...	45 investments ...
6 loudwater ...	16 hedge ...	26 jpmorgan ...	36 yoursection ...	46 bancorp ...
7 newsquest ...	17 mortgage ...	27 advisors ...	37 ex-dividend ...	47 stake ...
8 stocks ...	18 debt-to-equity ...	28 subsidiary ...	38 mill ...	48 tax ...
9 holdings ...	19 ltd ...			49 acquires ...
10 privilege ...	20 plc ...			50 wochit ...

Zeilen pro Seite: 50 1-50 of 100 1 / 2

Word (lowercase)	Frequency per million?	
	Focus	Reference
1 municipality	107.30	0.76 ...
2 railroad	63.60	0.66 ...
3 ncaa	43.40	0.19 ...
4 genus	90.89	1.48 ...
5 ohio	71.07	1.25 ...
6 pennsylvania	80.10	1.56 ...
7 illinois	64.56	1.23 ...
8 michigan	72.81	1.59 ...
9 missouri	43.40	0.61 ...
10 baseball	87.21	2.34 ...
11 kansas	44.06	0.72 ...
12 billboard	48.95	0.91 ...
13 voivodeship	25.60	0.02 ...
14 carolina	78.31	2.06 ...
15 gmna	24.80	0.00 ...
16 wisconsin	42.29	0.70 ...
17 nfl	45.96	0.85 ...
18 touchdown	31.37	0.29 ...
19 texas	118.54	3.91 ...
20 tennessee	41.85	0.84 ...
21 minnesota	44.23	0.96 ...
22 province	138.31	5.06 ...
23 ontario	55.84	1.48 ...
24 oklahoma	33.91	0.54 ...
25 criticized	34.58	0.57 ...

→ [https://cql.wortschatz-leipzig.de/#keywords?corpname=eng_news_2023_topic&tab=advanced&k_attr=word&k_itemsPerPage=50&ktab=keywords& ...](https://cql.wortschatz-leipzig.de/#keywords?corpname=eng_news_2023_topic&tab=advanced&k_attr=word&k_itemsPerPage=50&ktab=keywords&...)



Keyword-in-Context / Concordance

KONKORDANZ Eingelogg als private

business × CQL [lc="inflation"] • 551.824
396,26 freq. / m • 0.0094%

Linker Kontext KWIC Rechter Kontext

ID	Doc ID	Text
1	doc#56	2023-0... a broad-based and sharper-than-expected slowdown, with inflation higher than seen in<named_entity:DATE>several decades-
2	doc#117	2023-... solve the serious problems that face us from the border, to inflation , to the economy.</s><s>These are issues that people care
3	doc#166	2023-... ongress' subsequent passage of<named_entity:ORG>the Inflation Reduction Act</named_entity>in<named_entity:DATE>Nov
4	doc#218	2023-... sures to stabilize both the peso exchange rate and overall inflation .</s><s>He said that in<named_entity:DATE>November</i
5	doc#247	2023-... ar</named_entity>, but thanks to<named_entity:LAW>the Inflation Reduction Act</named_entity>, we were able to extend the
6	doc#247	2023-... ost-saving measures because of<named_entity:LAW>the Inflation Reduction Act</named_entity>.</s><s>That includes senic
7	doc#247	2023-... l_entity>a rebate if they try to raise their prices faster than inflation for drugs administered at a doctor's office.</s><s>"We're n
8	doc#267	2023-... med_entity:DATE>2022</named_entity>, including rising inflation rates and the depegging of TerraUSD,<named_entity:PER:
9	doc#269	2023-... n hopelessness.</s><s>Poverty is ravaging, hunger, fear, inflation , insecurity and the alarming rate of unemployment are my

Document metadata
(date, source, id, ...)

Keyword with
left/right context

→ https://cql.wortschatz-leipzig.de/#concordance?corpname=eng_news_2023_topic&tab=advanced&queryselector=cql&cql=%5Blc%3D%22inflation%22%5D&usesubcorp=business&showre

sults=1

<https://www.text-plus.org>



Statistics / Correlations

- Correlations to word *“inflation”* over time, by topic or source

	Domain	Frequenz
1	<input type="checkbox"/> seekingalpha.com	29.832 ...
2	<input type="checkbox"/> forextv.com	22.875 ...
3	<input type="checkbox"/> dailymail.co.uk	21.377 ...
4	<input type="checkbox"/> forexlive.com	17.775 ...
5	<input type="checkbox"/> marketscreener.com	8.637 ...
6	<input type="checkbox"/> businessmirror.com.ph	8.574 ...
7	<input type="checkbox"/> business.inquirer.net	8.110 ...
8	<input type="checkbox"/> biztoc.com	7.619 ...

	Doc.dayOfWeek	Frequenz
1	<input type="checkbox"/> Wednesday	111.557 ...
2	<input type="checkbox"/> Thursday	103.073 ...
3	<input type="checkbox"/> Tuesday	88.723 ...
4	<input type="checkbox"/> Friday	77.314 ...
5	<input type="checkbox"/> Sunday	61.097 ...
6	<input type="checkbox"/> Monday	60.128 ...
7	<input type="checkbox"/> Saturday	49.932 ...

	Doc.month	Frequenz ↓
1	<input type="checkbox"/> January	80.121 ...
2	<input type="checkbox"/> February	67.961 ...
3	<input type="checkbox"/> March	63.825 ...
4	<input type="checkbox"/> June	57.068 ...
5	<input type="checkbox"/> May	54.878 ...
6	<input type="checkbox"/> April	52.990 ...
7	<input type="checkbox"/> August	45.315 ...
8	<input type="checkbox"/> September	...
9	<input type="checkbox"/> July	...
10	<input type="checkbox"/> October	...
11	<input type="checkbox"/> November	20.124 ...

Work on subset with filter on certain month

	Topic classification using AGNews topics	Frequenz
1	<input type="checkbox"/> Business	551.824 ...

→ https://cql.wortschatz-leipzig.de/#concordance?corpname=eng_news_2023_topic&tab=advanced&queryselector=cql&attrs=word&viewmode=kwic&attr_allpos=all&refs_up=0&shorten_refs=1&...



Collocations

	Word	Kookkurrenzen ?	Kandidaten ?	T-score	MI	Log likelihood	LogDice ↓
1	✓ CPI	26.322	67.167	162,20	12,03	399.501,51	10,44 ...
2	✓ Calculator	20.990	60.701	144,84	11,85	311.814,80	10,13 ...
3	✓ Industry	20.942	144.996	144,62	10,59	269.470,98	9,94 ...
4	✓ Average	21.037	161.451	144,94	10,44	266.039,94	9,92 ...
5	<input type="checkbox"/> rate	28.260	467.262	167,85	9,33	312.320,25	9,83 ...
6	<input type="checkbox"/> high	28.922	561.491	169,75	9,10	310.130,69	9,73 ...
7	<input type="checkbox"/> Reduction	12.311	18.721	110,94	12,77	204.536,34	9,47 ...
8	<input type="checkbox"/> Act	12.100	109.830	109,91	10,20	148.497,45	9,23 ...
9	<input type="checkbox"/> rates	13.913	423.109	117,62	8,45	136.074,63	8,87 ...
10	<input type="checkbox"/> food	14.654	495.677	120,67	8,30	140.182,77	8,84 ...
11	<input type="checkbox"/> rising	9.287	143.245	96,23	9,43	103.642,30	8,77 ...
12	<input type="checkbox"/> down	15.736	705.025	124,92	7,89	141.634,28	8,68 ...
13	<input type="checkbox"/> data	9.750	271.288	98,48	8,58	97.043,89	8,60 ...
14	<input type="checkbox"/> Terms	20.940	1.237.272	143,90	7,49	177.024,53	8,58 ...
15	<input type="checkbox"/> halve	6.487	8.796	80,53	12,94	110.257,50	8,57 ...
16	<input type="checkbox"/> flight	6.536	79.780	80,75	9,77	76.080,88	8,41 ...
17	<input type="checkbox"/> core						
18	<input type="checkbox"/> higher	8.883	346.155	93,90	8,09	82.341,30	8,34 ...
19	<input type="checkbox"/> interest	9.417	415.656	96,64	7,91	84.934,00	8,32 ...
20	<input type="checkbox"/> bring	6.922	193.174	82,98	8,58	68.818,73	8,25 ...

Attribut ?
word

Skala ?
-5 -4 -3 -2 -1 KWIC 1 2 3 4 5

Custom range

Left and right context words for collocations

Funktionen anzeigen ?

- T-score
- MI
- MI3
- log likelihood
- min. sensitivity
- logDice
- MI.log_f

Eingelogg als private

KONKORDANZ

eng_news_2023_topic

business x CQL [l2="inflation"] • 551.824 396,26 freq. / m • 0.0094%

Filter CPI, Calculator, Industry, Average, Reduction, Act -3, 3 • 39.100 28,08 freq. / m • 0.00067%

Details

Linker Kontext KWIC Rechter Kontext

Examples for word selection

... ongress' subsequent passage of<named_entity:ORG>the **Inflation Reduction** Act</named_entity>in<named_entity:DATE>Nc

... ar</named_entity>, but thanks to<named_entity:LAW>the **Inflation Reduction** Act</named_entity>, we were able to extend th

... ost-saving measures because of<named_entity:LAW>the **Inflation Reduction** Act</named_entity>.</s><s>That includes seni

...<s>The headline<named_entity:ORG>Consumer Price **Inflation** </named_entity>(<named_entity> CPI) figure for<named_entity:DATE>the

...<s><s>Glossary of Retirement **Industry** Terms</s><s>CPI **Inflation** Calculator</s><s>CPI Average Price Calculator</s><s>Sr

6 doc#744 • 2023-...<s><s>Glossary of Retirement **Industry** Terms</s><s>CPI **Inflation** Calculator</s><s>CPI Average Price Calculator</s><s><n

8 doc#744 • 2023-...<s><s>Glossary of Retirement **Industry** Terms</s><s>CPI **Inflation** Calculator</s><s>CPI Average Price Calculator</s><s>Lo

8 doc#796 • 2023-...<s><s>Glossary of Retirement **Industry** Terms</s><s>CPI **Inflation** Calculator</s><s>CPI Average Price Calculator</s><s>Sr

9 doc#796 • 2023-...<s><s>Glossary of Retirement **Industry** Terms</s><s>CPI **Inflation** Calculator</s><s>CPI Average Price Calculator</s><s>Bit

10 doc#796 • 2023-...<s><s>Glossary of Retirement **Industry** Terms</s><s>CPI **Inflation** Calculator</s><s>CPI Average Price Calculator</s><s>Lo

11 doc#824 • 2023-...<s><s>Glossary of Retirement **Industry** Terms</s><s>CPI **Inflation** Calculator</s><s>CPI Average Price Calculator</s><s>Sr

→ https://cql.wortschatz-leipzig.de/#concordance?corpname=eng_news_2023_topic&tab=advanced&queryselector=cql&attrs=word&viewmode=kwic&attr_allpos=all&refs_up=0&shorten_refs=1&...



Pattern search / Token Information

RESULT DETAILS

CQL [lemma="inflation"] [pos_ud17="VERB" | pos_ud17="ADV" | pos_ud17="ADJ"] []{,4} [shape="Xxxx"] within <s/>

Number of hits	6
Number of hits per million tokens	0,28
Percent of whole corpus	0.0000281
Number of tokens (tokens)	21.320.935

Visual query
(CQL) builder

The CQL builder interface shows the query: [lemma="inflation"] [pos_ud17="VERB" | pos_ud17="ADV" | pos_ud17="ADJ"] []{,4} [shape="Xxxx"] within <s/>. It includes a visual representation of the query components and a result example section.

KONKORDANZ

eng_news_2023_topic2

CQL [lemma="inflation"] [pos_ud17="VERB" | pos_ud17="..."] • 6
0,28 freq. / m • 0.000028%

	Details	Linker Kontext	KWIC	Rechter Kontext
1	<input type="checkbox"/>	doc#2676 ->Top Stories at Middy:	Inflation Cools Down NOUN/inflation VERB/cool PROP/Down	; <named_entity:DATE> Jun
2	<input type="checkbox"/>	doc#2676 ->Top Stories at Middy:	Inflation Cools Down ; <named_entity:DATE> June NOUN/inflation VERB/cool PROP/Down PUNCT; PROP/June	</named_entity>Michigan
3	<input type="checkbox"/>	doc#4087 'named_entity'</s><s>	Inflation hit <named_entity:CARDINAL> 9.1% </named_entity> in <named_entity:DATE> June NOUN/inflation VERB/hit NOUN/9.1% ADP/in PROP/June	2022</named_entity>, its f
4	<input type="checkbox"/>	doc#8480 boost.</s><s>Headline	inflation dropped to <named_entity:CARDINAL> 6.8% </named_entity> in <named_entity:DATE> July NOUN/inflation VERB/drop ADP/to PROP/6.8% ADP/in PROP/July	</named_entity>, energy c
5	<input type="checkbox"/>	doc#8882 Annual</named_entity>	inflation slowed to <named_entity:CARDINAL> 3.3% </named_entity> in <named_entity:DATE> July NOUN/inflation VERB/slow ADP/to NOUN/3.3% ADP/in PROP/July	</named_entity>from from
6	<input type="checkbox"/>	doc#9391 Annual</named_entity>	inflation slowed to <named_entity:CARDINAL> 3.3% </named_entity> in <named_entity:DATE> July NOUN/inflation VERB/slow ADP/to NOUN/3.3% ADP/in PROP/July	</named_entity>from<nam

→ https://cql.wortschatz-leipzig.de/#concordance?corpname=eng_news_2023_topic2&tab=advanced&queryselector=cql&attrs=word&viewmode=kwic&attr_allpos=all&refs_up=0&shorten_refs=1&...





CQL <named_entity type=="ORG"/> [pos_ud17 != "CCONJ" & pos_ud17 != "PUNCT"]{1,8} <named_entity type=="ORG"/> within <s/>

Texttypen 4 (6) ...

CQL <named_entity type=="ORG"/> [pos_ud17 != "CCONJ" & pos_ud17 != "PUNCT"]{1,8} • 6.931

Sortieren GDEX x

325,08 freq. / m • 0.033%



Details

Satz

GDEX score

doc#13228 • 202...	<s><named_entity:PRODUCT>Bankman-Fried</named_entity>also testified that he doesn't recall ever directing<named_entity:ORG> Alameda </named_entity> employees not to spend the <named_entity:ORG> FTX </named_entity>customer deposits.</s>	0.899
doc#4496 • 2023...	<s><named_entity:ORG> State Street Corp </named_entity> lifted its stake in shares of <named_entity:ORG> NIKE </named_entity>by<named_entity:CARDINAL>1.3%</named_entity>during<named_entity:DATE>the 3rd quarter</named_entity>.</s>	0.85
doc#8237 • 2023...	<s><named_entity:ORG> Vanguard Group Inc. </named_entity> raised its stake in <named_entity:ORG> CrowdStrike </named_entity>by<named_entity:CARDINAL>2.0% </named_entity>in<named_entity:DATE>the first quarter</named_entity>.</s>	0.85
doc#10969 • 202...	<s>It was financed with a loan from<named_entity:ORG> the China Development Bank </named_entity> for <named_entity:ORG> 75% </named_entity>of the cost.</s>	0.829
doc#2335 • 2023	<s>Ellsworth Advisors<named_entity:ORG> LLC </named_entity> increased its stake	0.826

Relationship between two "ORG" NEs

→ https://cql.wortschatz-leipzig.de/#concordance?corpname=eng_news_2023_topic2&tab=advanced&queryselector=cql&attrs=word%2Cpos_ud17%2Clemma%2Cpos&viewmode=sen&attr_allpos=kw ...

Text+

<https://www.text-plus.org>



“Christmas market”

Not logged in
Log in



Text types 2 (2) ...

CQL [lemma="christmas"] [word="market"] [lemma="in" | le... • 45

Sort GDEX x

0.01 per million tokens • 7.7e-7%



Details

sentence

1	<input type="checkbox"/>	<input type="info"/>	2023-11-04 • ke...	<s>The<named_entity:DATE>annual</named_entity><named_entity:DATE> Christmas</named_entity>market in<named_entity:GPE>Canterbury </named_entity>city centre will open<named_entity:DATE>next weekend</named_entity>.</s>	<input type="copy"/>
2	<input type="checkbox"/>	<input type="info"/>	2023-11-04 • cb...	<s>She is now planning to sell her items at another<named_entity:DATE> Christmas</named_entity>market in<named_entity:GPE>Elmira </named_entity>.</s>	<input type="copy"/>
3	<input type="checkbox"/>	<input type="info"/>	2023-01-11 • in...	<s>She said she also rides the train for pleasure, for instance to visit the<named_entity:DATE> Christmas</named_entity>market in<named_entity:GPE>Wroclaw </named_entity>.</s>	<input type="copy"/>
4	<input type="checkbox"/>	<input type="info"/>	2023-10-03 • ne...	<s>The opening of<named_entity:GPE>Tallinn</named_entity><named_entity:DATE> Christmas</named_entity>market on<named_entity:DATE>November 25, 2022 </named_entity>.</s>	<input type="copy"/>
5	<input type="checkbox"/>	<input type="info"/>	2023-11-12 • ex...	<s><named_entity:CARDINAL>One</named_entity> reviewer said: “How is this supposedly the best<named_entity:DATE> Christmas</named_entity>market in<named_entity:LOC>Europe </named_entity>?</s>	<input type="copy"/>
6	<input type="checkbox"/>	<input type="info"/>	2023-11-18 • ed...	<s><named_entity:CARDINAL>Thousands</named_entity>of people flocked to the opening of Edinburgh's<named_entity:DATE> Christmas</named_entity>market on<named_entity:DATE>Friday </named_entity> <named_entity:TIME>evening</named_entity>.</s>	<input type="copy"/>

→ https://cql.wortschatz-leipzig.de/#concordance?corpname=eng_news_2023_topic&tab=advanced&queryselector=cql&keyword=war&attrs=word&viewmode=sen&attr_allpos=all&refs_up=0&...



Use-Cases and Cooperations



Wortschatz Leipzig – User Groups

Typical **user groups** include:

- Philologists, linguists, computer linguists
- Researchers in social science and other text-oriented domains
- NLP & digital humanities
- Translators, authors, journalists, ...
- General public interested in text generation
- ...



WestAI



- One of four German “AI service centers” (*Center West*), funded by the **BMBF**.
- Important focus: the creation and provision of **large, multimodal AI models**.
 - Delivery of a **large collection of high-quality Wortschatz corpora** from 2021-2022 to WestAI and its partnering institutes and projects.
 - The data are to be used for text mining and the development of **German large language models**.

<https://westai.de>



LLM Training at ScaDS.AI



- *Center for Scalable Data Analytics and Artificial Intelligence Dresden/Leipzig*
- Collaboration for a research project to train a state-of-the-art **Large Language Model (LLM) for German**
- The *Wortschatz Leipzig project* can deliver the **large amount of raw text** necessary for training.
- In addition, personnel can lend their **expertise** in **web crawling** and **text/linguistic preprocessing**



Project WINTER



- **EU-funded project**, which aims to develop a web interactive platform for the management of coal regions in Germany, Greece and Poland.
- **26 Wortschatz news corpora for German, Greek and Polish** were used by the TH Georg Agricola (THGA) for an analysis of **media representation and perception** of the structural reduction of coal industry in the different countries.
- Utilized the **Wortschatz portal** as an intuitive, easily usable tool.

<https://winter-project.eu>



Sentiment analysis on customer reviews

- Extensive stock of product reviews as a corpus for identifying customers' sentiment about various product characteristics
- *SentiWS* containing 34K German words with their sentiment polarity $[-1,+1]$ (German language only) combined with word embeddings for a larger recall
- Identifying
 - Positive/negative characteristics of single products
 - Summary of negative characteristics of product groups
 - Positive/negative characteristics of products in comparison with their product groups

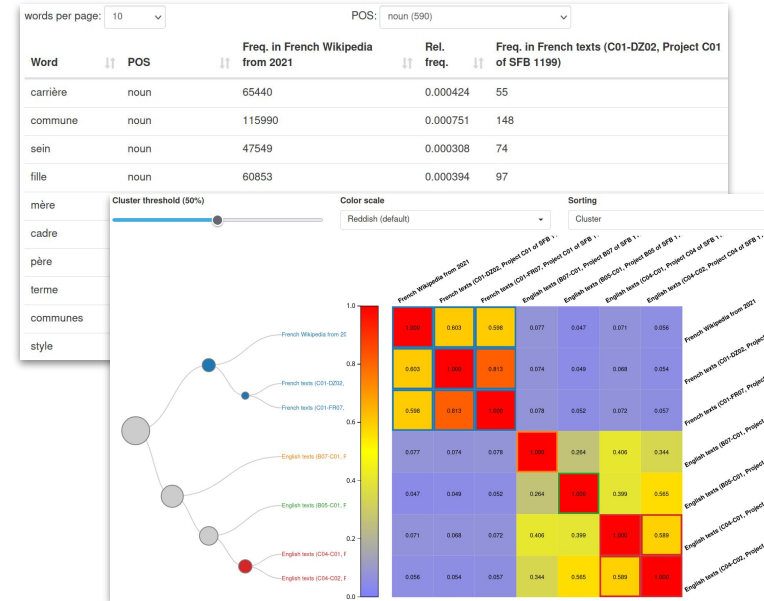
→ <https://wortschatz-leipzig.de/en/download#sentiWSDownload>



SFB 1199 “Processes of Spatialization under the Global Condition”



- Processing of historical text data
 - Texts of geographical societies of three countries from the 19th and 20th centuries
 - Collections of magazines and books from the 20th century in different languages
 - Pre-revolutionary French dictionaries
- Corpus analysis and terminology extraction
- Approach: distant reading combined with close reading “drill-down”
- Analysis of vocabulary change



Other users

... include **private corporations**,



... **researchers**, from a multitude of institutions

... and **you?**





WORTSCHATZ LEIPZIG

<https://wortschatz-leipzig.de>

